

معرفی کتاب: مبانی علوم کامپیوتر جدید

عباس محرابیان

دانشگاه بریتیش کلمبیا

abbas.mehrabian@gmail.com

۹ خرداد ۱۳۹۵

علوم کامپیوتر شاخه‌ای بسیار پویا و در حال تغییر است. فراگیر شدن اینترنت، شبکه‌های اجتماعی و موبایل‌ها مسائل جدیدی به وجود آورده‌اند که حل آن‌ها با روش‌های سنتی ممکن نیست. سه دانشمند برجسته علوم کامپیوتر در حال نوشتن کتابی در زمینه مبانی علوم کامپیوتر جدید هستند. یکی از ویژگی‌های دوران جدید اینست که می‌توان به (نسخه غیرنهایی) یک کتاب پیش از انتشار آن دسترسی داشت^۱. خواندن این کتاب یا دست‌کم نگاه کردن به فهرست مطالبش را به همه علاقمندان به علوم کامپیوتر، هم علاقمندان به مباحث نظری و هم عملی، اکیداً پیشنهاد می‌کنم. در ادامه ترجمه‌ای از مقدمه کتاب آمده است.

«علوم کامپیوتر به عنوان یک رشته دانشگاهی در دهه ۶۰ میلادی متولد شد. مباحث اصلی در آن هنگام زبان‌های برنامه‌نویسی، کامپایلرها، سیستم‌های عامل و ریاضیات موردنیاز برای این زمینه‌ها بودند، و لذا درس‌هایی مثل نظریه زبان‌ها و ماشین‌ها و پیچیدگی محاسبه در دانشگاه‌ها تدریس می‌شدند. در دهه ۷۰، نظریه الگوریتم‌ها به عنوان یک مبحث کلیدی مطرح شد. تأکید روی «کامپیوتر» بود و این که چطور می‌توان از آن بهتر استفاده کرد.

در دهه‌های اخیر تغییراتی اساسی رخ داده و تأکید بیشتری روی داده^۲ هست و این که چطور می‌توان آن را معنا کرد و از آن بهتر استفاده کرد. علت اینست که امروزه ما فناوری‌های لازم برای جمع‌آوری و نگهداری حجم‌های وسیعی از داده را پیدا کرده‌ایم (مثلاً داده‌های مربوط به ژن‌های انسان‌ها، تراکنش‌های مالی، و شبکه‌های اجتماعی مجازی). به علاوه، روشن شده است که با تحلیل درست داده‌ها می‌توان به نتایج شگفت‌آوری رسید. یک نمونه جالب، برنامه کامپیوتری است که شرکت گوگل برای بازی GO طراحی کرده است. این برنامه داده‌های مربوط به ۱۰۰ هزار بازی واقعی را تحلیل کرده و

^۱ نسخه ژانویه ۲۰۱۶ کتاب از صفحه شخصی یکی از نویسندگان قابل دریافت است:

www.cs.cornell.edu/jeh/bookJan25_2016.pdf
data^۲

دو ماه پیش موفق شد قهرمان جهان را شکست بدهد.^۳

بنابراین استفاده از کامپیوتر برای فهم و تحلیل و استفاده مناسب از داده، مسئله جدیدی است که برای پژوهش‌گران علوم کامپیوتر مطرح شده است. هدف از نوشتن این کتاب این بوده که مبانی نظری لازم برای چنین پژوهش‌گرانی را پوشش دهیم، و نویسندگان تلاش کرده‌اند ریاضیاتی که در این کتاب آمده ریاضیاتی باشد که طی ۴۰ سال آینده به درد محققین علوم کامپیوتر خواهد خورد. یک تغییر عمده نسبت به گذشته، تغییر تمرکز از ریاضیات گسسته به آمار و احتمال، جبر خطی و روش‌های عددی می‌باشد.

نسخه‌های اولیه این کتاب برای تدریس در دانشگاه‌های معتبری مثل CMU و کورنل در مقاطع کارشناسی و کارشناسی ارشد استفاده شده است. پیش‌نیازهای ریاضی لازم در پیوست کتاب آمده‌اند. شایان ذکر است که تأکید کتاب بر ارائه مبانی ریاضی است و نه بررسی عمیق یک یا چند مسئله کاربردی.

کتاب با بحث درباره هندسه ابعاد بالا آغاز می‌شود. امروزه در شاخه‌هایی مثل جستجو و بازیابی اطلاعات و یادگیری ماشین، به داده‌ها عموماً به صورت بردارهایی با تعداد زیادی مؤلفه نگاه می‌شود. نمایش این چنینی اگرچه در برخی موارد چندان طبیعی نیست، معلوم شده از نظر عملی مزیت‌های متعددی دارد و الآن دیگر استاندارد شده است. شهود ما درباره فضاهای دو و سه بُعدی می‌تواند در فهم فضاهای ابعاد بالاتر بسیار گمراه‌کننده باشد. در فصل دوم کتاب به تفاوت‌های اصلی بین ابعاد پایین و بالا می‌پردازیم.

برای کار کردن با داده‌ها در ابعاد بالا، شناخت ماتریس‌ها و کار کردن با آن‌ها اهمیت به‌سزایی دارد. یک قضیه مرکزی درباره ماتریس‌ها، *singular value decomposition* است که کاربردهای نظری و عملی متعددی دارد. در فصل چهارم به جنبه‌های الگوریتمی این قضیه و کاربردهایش می‌پردازیم.

برای فهم بهتر شبکه‌های بزرگی مثل اینترنت و شبکه‌های اجتماعی، باید مدل‌های ساده‌ای داشته باشیم که تا حد ممکن شبیه این شبکه‌ها باشند. ما ساده‌ترین مدل این چنینی را بررسی می‌کنیم، که گراف تصادفی Erdős-Rényi نام دارد. این بررسی نشان می‌دهد که چطوری یک سری رفتارهای جزئی مربوط به تک‌تک رئوس می‌تواند منجر به پدیده‌های فراگیر مثل مؤلفه‌های همبندی بزرگ بشود. چند مدل دیگر از گراف‌های تصادفی هم در این کتاب بررسی می‌شوند.

یکی از مشاهدات بسیار جالب در علوم کامپیوتر در دو دهه اخیر این بوده که یک سری روش‌های کلی توانسته‌اند مسائلی در زمینه‌های بسیار متفاوتی را حل کنند. یک مثال عمده یادگیری ماشین^۴ بوده که مسائلی در زمینه بینایی ماشین، سیستم‌های پیشنهاددهنده، و تحلیل زبان‌های طبیعی را توانسته حل کند. ما مبانی یادگیری ماشین را مرور می‌کنیم و مثلاً پارامتر VC-dimension را معرفی می‌کنیم، که دقیقاً به ما می‌گوید برای «یادگیری» به چه مقدار داده نیاز داریم. یک روش کلی دیگر استفاده از زنجیرهای مارکف و قدم‌زدن تصادفی است. ریاضیات مربوطه و ارتباطات آن با شبکه‌های الکتریکی را هم مرور می‌کنیم.

در نظریه کلاسیک الگوریتم‌ها فرض بر اینست که ورودی مسئله در حافظه قرار دارد و الگوریتم همواره به آن دسترسی

^۳ <http://goo.gl/eWLH2F>
^۴ machine learning

دارد. این فرض در مورد بسیاری از مسائل جدید صادق نیست. چندین مدل محاسباتی جدید برای کار با این مسائل مطرح شده‌اند، مثل مدل‌های online و streaming که در فصل هفتم بررسی می‌شوند. یک ایده مرکزی برای کار با این مدل‌ها گلچین کردن^۵ است. در این فصل بحث می‌کنیم چطور می‌توان خوب داده‌ها را گلچین کرد و چه پارامترهای آماری و جبرخطی‌ای را با چه دقتی می‌توان تخمین زد.

یک ابزار عمده مدرن برای کار با داده، تقسیم‌بندی آن به گروه‌هایی کوچک‌تر است به طوری که اعضای هر گروه شبیه هم‌اند.^۶ در این کتاب برخی روش‌های کلاسیک برای گروه‌بندی (مثل k-means) و هم‌چنین الگوریتم‌های جدیدتر را معرفی می‌کنیم. به علاوه در مورد ارزیابی کیفیت گروه‌بندی هم بحث خواهیم کرد.

دیگر مباحثی که در کتاب مطرح می‌شوند مدل‌های گرافی^۷، روش‌های مبتنی بر پخش شایعات^۸، روش‌های رتبه‌بندی و رأی‌گیری، و نهایتاً بازیابی اطلاعات از دست رفته^۹ هستند. در پیوست کتاب مجموعه پیش‌نیازهای ریاضی جمع‌آوری شده‌اند.»

sampling^۵
clustering^۶
graphical models^۷
belief propagation^۸
compressed sensing^۹